On the identification of the nature of behavioural dependence in two-sample capture-recapture study

Kiranmoy Chatterjee*

Diganta Mukherjee †

Abstract

Motivated by various potential applications, problem of estimating human population size from dependent dual-record system (DRS) is a very challenging task. Owing to the non-identifiability of otherwise appropriate M_{tb} model under DRS, our contribution lies in the construction of some strategies for classifying the nature of underlying behavioural dependency of the individuals belonging to a given population (i.e. whether the population is *recapture prone* or *averse*). This classification strategy would be quite appealing in improving the inference as evident from recent literature. Simulation studies and application on three different real life data sets are carried out to explore the performance of this strategy.

*Department of Statistics, Bidhannagar College, Kolkata, India and Indian Statistical Institute, Kolkata, India; E-mail: *kiranmoy07@gmail.com*

†Indian Statistical Institute, Kolkata, India.

Key words: Classification; Directional nature of behavioural response effect; Human population; Randomized rule; Recapture probability.

1 Introduction and Motivation

Estimation of size of a given population is an important statistical issue which has vast application in the field of Government statistics, demography, epidemiology and animal abundance. In practice, it is mostly impossible to count all the individuals in the population accurately by a census, specially when population is large enough and/or very hard to reach the individuals. As a remedy, more than one attempt is carried out independently, near to the census operation, and the population size (say, N) is estimated by matching the available lists (two or more) of information. This kind of data structure by matching lists is known as Multiple-record system and this is equivalent to the capture-recapture system, popularly relevant in biological studies. In the context of closed human population, more than two sources of information is hardly found. When two attempts has been made to obtain N in capture-recapture format, then such data structure is known as Dual-record System (DRS) or simply, Dual System. Coverage error estimation in census or vital events, estimation of epidemiological surveillance, estimation of the size of different hard-to-count populations are the primary applications for human population. Possible heterogeneity in capture probabilities among individuals in the population can be factored out with the proposal on formation of homogeneous post-strata given by Chandrasekar and Deming (1949), which is widely implemented in real life applications.

After constructing such mutually exclusive post-strata, which are within homogeneous

but between heterogeneous, relevant statistical models for DRS can be analysed for each of those post-strata. Model M_t (Otis et al., 1978), popularly known as Petersen Model (Wolter, 1986), has received much attention in practice among all the relevant models mainly because of its simplicity. If $x_{1\cdot}$ and $x_{\cdot 1}$ denote total number of individuals present in first and second lists respectively, x_{11} denotes the number of common persons present in both the lists (refer to Table 1 for DRS data structure) and x_0 denotes the total number of distinct individuals counted by the DRS, estimate of N from the model M_t would be $x_0 + \hat{x}_{00}$ or $\frac{x_{1\cdot}x_{\cdot 1}}{x_{11}}$, following the underlying assumption of *causal independence* in M_t. This estimator also popularly known as Petersen estimate or Chadrasekar-Deming estimate. More details on this model including various estimates can be found in Chatterjee and Mukherjee (2016a).

Insert Table 1

However, this assumption may seriously mislead in many situations for human population. Several methodologists and practitioners (*see* Chandrasekar and Deming, 1949; Greenfield, 1975; El-Khorazaty, 2000; Jervis et al., 2000) argued that the independence assumption may not be justifiable in reality. Often, this independence is violated due to the presence of behavioural response variation at the time of second capturing attempt in DRS. Often an individual who is captured by first attempt may have more chance to be included in the second list than the individual who has not been captured by first attempt. If it is, then the corresponding population is treated as *recapture prone*. This kind of behavioural connection at the time of second attempt is commonly encountered in demographic study.

Otherwise, for reverse case, the population becomes *recapture averse*. In case of hard-to-count populations, drug addicted populations etc., *recapture aversion* is often observed. These kind of changes in behaviour of an individual at the time of second time capturing may occur due to different causes (*see* Wolter, 1986) and this feature is grossly known as behavioural response variation. When both the time variation effect (*t*) and behaviour response variation effect (*b*) acts together, model M_{tb} can be treated as the most general and relevant statistical model for capture-recapture data under homogeneity. Gosky and Ghosh (2011) showed the appropriateness of this model among all the capture-recapture models proposed in Otis et al. (1978).

Now we state some basic notation before proceeding further. Let p_{ij} be the probability attached to each individual to be included in the count x_{ij} in (i, j)th cell of Table 1, where $i, j \in \{0, 1, \cdot\}$. In addition we denote

$$\begin{aligned} Pr(\text{An individual is captured in List-2} \mid \text{S/He} \in \text{List-1}) &= \frac{p_{11}}{p_{1\cdot}} = c, \\ Pr(\text{An individual is captured in List-2} \mid \text{S/He not} \in \text{List-1}) &= \frac{p_{01}}{1 - p_{1\cdot}} = p. \end{aligned}$$

We consider $c \neq p$ which refers a violation of *causal independence*. Thus, there always exists some constant $\phi(> 0)$ such that $c = \phi p$. This ϕ is termed as *behavioural response effect*. Thus alternatively, $\phi > 1$ (equivalently, $c > p$) refers to positive association between the two sources and associated population is said to be *recapture prone* and $\phi < 1$ (equivalently, $c < p$) refers to negative association and associated population is said to be *recapture averse*. Hence, the newly advanced model incorporating the non-existence of *causal independence* is denoted as M_{tb}. Therefore, model M_{tb} suffers from a estimability problem as ϕ (or p) is not estimable but the product $\phi p = c$ is estimable. The likelihood

for model M_{tb} fails in DRS. Thus, we have a model M_{tb}, that makes intuitive sense of real scenario but is not, however, identifiable. A complete account of the model M_{tb} in DRS can be found in Chatterjee and Mukherjee (2016b).

However, some relevant knowledge on the uncertainty of ϕ, if available, might help in drawing reasonably good inference on N, as evident from literature (*see* Nour, 1982; Chatterjee and Mukherjee (2016b); Chatterjee, 2016). Particularly if underlying ϕ for any population is correctly known to be greater than 1 (or less than 1), then uncertainty on ϕ will be reduced since the domain of ϕ shrinks to $(1, \infty)$ (or $(c, 1)$), where c is the recapture probability of an individual at the time of second time capturing occasion. Hence, one can expect that inference would be better if that available knowledge is used. This issue has been proved empirically in Chatterjee and Mukherjee (2016b) and Chatterjee (2016). Intuitively, one may suggest that a given population is *recapture prone* if $\hat{c}$ is very close to 1. On the other hand, if it is close to 0, then associated population would be *recapture averse* with high probability. But giving idea about the possible direction of ϕ is always a challenging job if $\hat{c}$ is neither close to 1, nor close to 0. As per our knowledge, no strategy has been developed yet to understand the directional nature of ϕ from the given data only. Therefore, in this article, our aim is to develop some competing classification strategies in order to infer about the underlying directional nature of behavioral dependence of a given population, i.e. whether the population is *recapture prone* or *averse* or *causally independent*.

In the next section we formulate three classification strategies and therefore, in section 3, we illustrate our proposed strategies and evaluate their performance through an extensive simulation study. In section 4 we apply our proposed strategies in order to classify the

underlying dependency nature of the populations in three real datasets comprising three different fields of application. Finally, in concluding section, we discuss the implication, advantages and possible extension of the proposed classification strategies.

2 Proposed Classification Strategies

Nour(1982) proposed a strategy for estimating the unobserved cell x_{00} in Table 1 based on some assumptions.

In this section, we formulate a strategy for classification of the population in terms of the directional nature of ϕ, following the idea of Nour(1982) but with only one mild assumption that the conditional probability $p \geq 1/3$. This assumption implies that $3c-\phi \geq 0$ as $c = p\phi$. The *mle* $\hat{c} = (x_{11}/x_{1\cdot})$ is consistent and efficient estimate of c. Hence, in terms of data we can approximately write that $\phi = \hat{c}p^{-1}$ for sufficiently large N. Therefore, as $N \geq x_0$,

$$\begin{aligned} \phi = \frac{x_{11}}{x_{1\cdot}}\frac{N - x_{1\cdot}}{x_{01}} &\geq \frac{x_{11}}{x_{1\cdot}} \\ or, (x_{1\cdot}\phi - x_{11}) &\geq 0. \end{aligned}$$

Thus, for constructing a classification strategy for ϕ, consider the inequality

$$\frac{(3\hat{c} - \phi)(x_{1\cdot}\phi - x_{11})}{x_{\cdot 1}\phi} \geq k,$$

where k is some nonnegative real number. Hence the above inequality may be expressed as

$$\phi^2 + \phi\tfrac{kx_{\cdot 1} - 4x_{11}}{x_{1\cdot}} + 3\hat{c}^2 \leq 0 \tag{1}$$

$$or, \quad (\phi - \phi_0)(\phi - \phi_1) \leq 0, \tag{2}$$

where ϕ_0 and ϕ_1 are two real roots of the quadratic equation (1) when equality holds and satisfy $\phi_0\phi_1 = 3\hat{c}^2$ and $\phi_0 + \phi_1 = (4x_{11} - kx_{10})/x_{1\cdot}$ and $\phi_0 < \phi < \phi_1$. Since, A.M. $\geqslant$

G.M. for the two roots ϕ_0 and ϕ_1, then

$$k \leqslant (4 - 2\sqrt{3})x_{11}/x_{\cdot 1},$$

equality holds only when $\phi_0 = \phi_1 = \phi = \sqrt{3}\hat{c}$. In addition, as $k \geq 0$ holds under the assumption that $p \geq 1/3$, which usually holds for human population, the values of ϕ_0 and ϕ_1 corresponding to this lower bound are $\hat{c}$ and $3\hat{c}$ respectively. Furthermore, the root ϕ_0 is a monotonically increasing function, while ϕ_1 is a monotonically decreasing function, of k. This implies $\hat{c} \leq \phi_0 \leq \sqrt{3}\hat{c}$ and $\sqrt{3}\hat{c} \leq \phi_1 \leq 3\hat{c}$.

Now, for given ϕ_0 and ϕ_1, $\exists$ some $\xi \in (0, 1)$ such that (2) may be written as $\phi = \xi\phi_1 + (1 - \xi)\phi_0$, which implies $\xi = (\phi\phi_0 - \phi_0^2)/(3\hat{c}^2 - \phi_0^2)$ using the relation $\phi_0\phi_1 = 3\hat{c}^2$. Considering ξ as a function of ϕ_0, it is noted that for given ϕ and data, ϕ_0^* is a point at which ξ is maximum, where

$$\phi_0^* = \frac{3\hat{c}^2}{\phi} - \frac{3\hat{c}^2}{\phi}\sqrt{1 - \frac{\phi^2}{3\hat{c}^2}}$$

and that the corresponding value of ξ is ξ^* such that

$$\xi^* = \frac{1}{2}\left(1 - \sqrt{1 - \frac{\phi^2}{3\hat{c}^2}}\right)$$

and $\phi_1^* = 3\hat{c}^2/\phi_0^*$.

Now by definition, the value $\phi_0 = \phi_0^*$ is a lower bound for ϕ, i.e. $\phi > \phi_0^*$ which implies $\phi < \sqrt{3}\hat{c}$. This is also a necessary condition for both ϕ_0^* and ξ^* to be real-valued.

It is in the nature of the current problem that additional information is required in order to obtain exact inference for ϕ. We make the additional assumption that for a fixed ϕ and a fixed $\hat{c}$, ϕ_0^* has the same range of values given to ϕ_0 under the present structure. Thus,

$$\hat{c} \leq \phi_0^* \leq \sqrt{3}\hat{c},$$

from which,

$$3\hat{c}/2 \leq \phi \leq \sqrt{3}\hat{c}. \tag{3}$$

Alternatively, the bound on ϕ_1^* (i.e. $\sqrt{3}\hat{c} \leq \phi_1^* \leq 3\hat{c}$) also leads to (3). Thus, we have a more tight bound for possible domain of ϕ under some mild assumptions. The assumption that ϕ_0^* has the same range of values as that assumed for ϕ_0, need to be validated. The inequality $\phi_0^* \leq \sqrt{3}\hat{c}$ is always valid since all ϕ_0's, including ϕ_0^*, have $\sqrt{3}\hat{c}$ as the maximum. On the other hand, the inequality $\phi_0^* \geq \hat{c}$ does not necessarily hold for all ϕ_0's.

Now we present three classification rules for inferring about the type of behavioural dependency for a given population. The following three rules have potential to identify when there is no or little evidence for causal dependence, i.e. when ϕ is either 1 or very close to 1. Let us consider that ϕ_l and ϕ_u respectively be the lower and upper tolerance limit of ϕ, for which one may say that underlying DRS model is causally independent. For example, if we consider 5% tolerance, then (ϕ_l, ϕ_u) = (0.95, 1.05).

Rule I. Taking cue from Nour (1982), it is proposed that the lower bound in (3), which results from setting $\phi_0^* = \hat{c}$, be taken as a threshold for suggesting the type of behavioural nature. So, if $3\hat{c}/2 > \phi_u$ we say the population is *recapture prone*. Again if $3\hat{c}/2 < \phi_l$, we say the population is *recapture averse* and thus, the population will be called as *causally independent* if $\phi_l < 3\hat{c}/2 < \phi_u$. We find that Nour's technique is rather conservative as it has a tendency towards indicating *recapture aversion*.

Rule II. Admitting the conservativeness of the previous classification rule, a second

rule is set with the mid-value of the range of ϕ in (3). If mean of the upper and lower limits in (3), i.e. $1.616\hat{c}$, is above ϕ_u, then we call the population *recapture prone*. On the contrary, it will be *recapture averse*, if the same is below ϕ_l. Finally, the population will be considered as *causally independent* if $\phi_l < 1.616\hat{c} < \phi_u$. This rule increases the chance of inferring a population as *recapture prone* and therefore, reduces the bias in ***Rule I***.

Rule III. Here we propose a randomized rule to identify the direction of the underlying behavioural dependency. We consider the following steps to infer about the behavioral classification of a given population.

Step 1: Carry out a randomized trial based on a Bernoulli r.v., say X_p, with the following probability function in favour of recapture proneness of the given population.

$$\psi_p(\hat{c}) = \begin{cases} 1 & \textit{if} \quad \frac{3\hat{c}}{2} > \phi_u \\ \delta_p & \textit{if} \quad \frac{3\hat{c}}{2} \leq \phi_u < \sqrt{3}\hat{c} \\ 0 & \textit{if} \quad \sqrt{3}\hat{c} \leq \phi_u, \end{cases},$$

where

$$\delta_p = \max\left\{0, 1 - \left(\phi_u - \frac{3\hat{c}}{2}\right) \Big/ \left(\sqrt{3}\hat{c} - \frac{3\hat{c}}{2}\right)\right\},$$

Step 2: If the given population is not found to be recapture prone in Step 1 i.e., if X_p is not observed to be 1, carry out another randomized trial based on another Bernoulli r.v.,

say, X_a with the following probability function in favour of recapture aversion.

$$\psi_a(\hat{c}) = \begin{cases} 1 & \textit{if} \quad \sqrt{3}\hat{c} < \phi_l \\ \delta_a & \textit{if} \quad \frac{3\hat{c}}{2} \leq \phi_l \leq \sqrt{3}\hat{c} \\ 0 & \textit{if} \quad \frac{3\hat{c}}{2} > \phi_l, \end{cases},$$

where

$$\delta_a = \left(\phi_l - \frac{3\hat{c}}{2}\right) \Big/ \left(\sqrt{3}\hat{c} - \frac{3\hat{c}}{2}\right).$$

Step 3: If the given population is not found to be recapture averse in Step 2 i.e., if X_a is not observed to be 1*, therefore, the given population is classified as causally independent.*

Thus, when the probability is not certain (i.e., 1 or 0), one has to perform a bernoulli experiment with probability of *recapture proneness* equal to δ_p and *recapture aversion* equal to δ_a, in order to classify whether a given population is *recapture prone* or *averse*.

The effective probabilities for considering an individual as recapture prone (RP) or recapture averse (RA) or causally independent (CI) are computed based on Prob($X_p = 1$), Prob($X_a = 1, X_p = 0$) and Prob($X_a = 0, X_p = 0$) respectively. Probabilities are computed based on the asymptotic normality of *mle* $\hat{c}$. These probabilities are presented in the following theorem and proof of the theorem is sketched in appendix.

Theorem 1. *Large sample effective probabilities for considering an individual to be re-capture prone or averse or causally independent defined in Rule III are as follows:*

$$Pr(RP) = 1 - \left[\delta_p F\left(\frac{\phi_u}{\sqrt{3}}\right) + (1-\delta_p)F\left(\frac{2\phi_u}{3}\right)\right],$$

$$\begin{aligned} Pr(RA) &= \delta_a F\left(\tfrac{2\phi_l}{3}\right) + (1-\delta_a)F\left(\tfrac{\phi_l}{\sqrt{3}}\right) \quad \textit{if} \quad \tfrac{2\phi_l}{3} < \tfrac{\phi_u}{\sqrt{3}}, \\ &\quad (1-\delta_a)F\left(\tfrac{\phi_l}{\sqrt{3}}\right) + \delta_a(1-\delta_p)F\left(\tfrac{2\phi_l}{3}\right) + \delta_a\delta_p F\left(\tfrac{\phi_u}{\sqrt{3}}\right) \quad \textit{if} \quad \tfrac{2\phi_l}{3} > \tfrac{\phi_u}{\sqrt{3}} \\ Pr(CI) &= \left[\delta_p F\left(\frac{\phi_u}{\sqrt{3}}\right) + (1-\delta_p)F\left(\frac{2\phi_u}{3}\right)\right] \\ &\quad - \left[\delta_a F\left(\tfrac{2\phi_l}{3}\right) + (1-\delta_a)F\left(\tfrac{\phi_l}{\sqrt{3}}\right)\right] \quad \textit{if} \quad \tfrac{2\phi_l}{3} < \tfrac{\phi_u}{\sqrt{3}}, \\ &\quad \left[\delta_p(1-\delta_a)F\left(\frac{\phi_u}{\sqrt{3}}\right) + (1-\delta_p)F\left(\frac{2\phi_u}{3}\right)\right] \\ &\quad - \left[\delta_a(1-\delta_p)F\left(\tfrac{2\phi_l}{3}\right) + (1-\delta_a)F\left(\tfrac{\phi_l}{\sqrt{3}}\right)\right] \quad \textit{if} \quad \tfrac{2\phi_l}{3} > \tfrac{\phi_u}{\sqrt{3}}, \end{aligned}$$

where $F(\cdot)$ *cumulative distribution function of std. normal variate* $[(\hat{c}-c)/s.d.(\hat{c})]$.

Theorem 1 is very useful in applied work in that it provides a quite important and reasonably simple empirical strategy for detecting behavioural dependence without any need for additional information. Note that for some configuration of $p'_{ij}s$ (as defined in the introduction), the boundary contraints in the probability calculations may become binding. In such cases, the conclusion from Theorem 1 would only be approximate in nature, apart from the usual normality approximation. In all other cases, the strategy will work well. Graphical comparative study between three effective probabilities calculated in Theorem 1 has been given in the next section.

3 Simulation Study

3.1 Evaluation of classification rules

We consider 12 simulated populations characterized by different pairs of capture probabilities ($p_{1\cdot}$, $p_{\cdot 1}$) that are presented in Table 2. Further, we also consider three possible situations of behavioural effect - (*i*) *recapture proneness* (*RP*) represented by $\phi = 1.50$, (*ii*) *recapture aversion* (*RA*) represented by $\phi = 0.60$ and (*iii*) *causal independence* (*CI*) refers $\phi = 1.00$. These 12 simulated populations for each of three said ϕ values together encompasses all possible combinations. 4 populations are chosen for each of the three absolute difference values, (0.1, 0.15, 0.20), between $p_{1\cdot}$ and $p_{\cdot 1}$. The true value of the parameter c is also presented for each population. We take the true population size (N) as 1000.

Insert Table 2

Table 3 presents the performance evaluation of the developed classification strategies in section 2, in terms of *correct classification rate* (CCR) of the underlying directional nature of ϕ. CCR is presented in percentage (%) after computing the number of correct classification out of 5000 replications for each simulated population.

Insert Table 3

Empirical evaluation of classification *Rule I* (columns 2, 6 & 10 in Table 3) shows that this classification strategy works efficiently except for the *recapture prone* populations P4, P8, P11 and P12 as well as for the *causally independent* populations I1, I2, I4, I6, I8, I10

and I12. Performance of the second strategy, presented in columns 3, 7 & 11 in Table 3 is also efficient, except for a lesser number of situations (i.e. P4, P8, P12) for causally dependent populations, where it fails. Indeed, *Rule II* produces improvement over *Rule I* towards the correct classification in almost all cases except for the truly independent populations. *Rule II* may identfy only a few causally independent populations, such as I1, I7 and I11. The reason behind such weak performance is due to smaller length of the interval for $\hat{c}$ falling into which one would opt the given population as causally independent as per *Rule II*. Further, it can be observed that both the two classification rules *I* & *II* fail for those *recapture prone* populations which represents too small recapture probabilities ($p_{\cdot 1}$). Lastly, more improvements have been established in columns 4, 8 & 12, as *Rule III* increases the rate of correct classification, specially for those situations where *Rules I* and *II* failed. By definition, incorporation of randomized decision (based on δ_p and δ_a) has increased the chance of detecting a population to be causally independent. These notable betterment helps us to make correct classification for more *causally independent* populations (*see* results for populations I1, I5, I7, I8, I9 & I11). The analysis of the performance of *Rule III* will be better understood in the next section.

3.2 Illustration of effective probabilities

Here we consider the same 12 duplets on the values of $(p_{1\cdot}, p_{1\cdot})$ for $N = 1000$ which are considered in Table 2. Comparative study on the effective probabilities calculated in Theorem 1 are presented in Figure 1. As one reads the panel from left to right, we see that the classification performance is quite satisfactory for the first three columns. The fourth

column (i.e. 4th, 8th and 12th pair of $(p_{1\cdot}, p_{\cdot 1})$) in the picture show erratic performance with classification being biased towards aversion or indifference. Recall that these are the configuration which had very low List 2 probability ($p_{\cdot 1}$) as well as low CCR in Table 3 and hence the results are consistent with our earlier intuition. Again, other *causally independent* populations, I2, I6, I10, having low CCR unanimously, actually posses very high List 2 probability. Overall, we can see that our theoretical intuition put forth in the discussion after Theorem 1 are also carried over here.

Insert Figure 1

4 Real Data Illustration

In this section we illustrate the three strategies proposed in section 2 through the application to the populations associated with three real datasets originated from three different fields - demographic, epidemiological and socio-economic surveys.

Firstly, we consider the **Malawi Death data** obtained from a Population Change Survey to estimate birth, death and migration rates conducted by the National Statistical Office in Malawi between 1970 and 1972. Greenfield (1975) introduced this dataset. Later, it has also been used by Nour (1982) and Chatterjee and Mukherjee (2016b). Very large value of $\hat{c}$ for all the strata, except *Lilongwe*, clearly indicate *recapture proneness*. Thus, we consider the data for *Lilongwe* and *Other Urban Areas* (*see* top panel of Table 4) for comparative analysis of the proposed classification rules.

Secondly, we use a **Homicide data**, which is analysed by Eckberg (2000) to establish

the utility of dual enumeration methods for estimating the total number of unrecorded murders occurred in South Carolina, 1877-1878. This interesting work was meant for tracing the historic trends in homicide based on the 2 sources alone and the author claimed that method with popular Chandrasekar-Deming estimate would face a formidable undercount problem. This happens due to a possible positive correlation between two data sources - (*i*) South Carolina Department of Archives and History and (*ii*) News and Courier Reports. Dual system data was available for total 33 counties in South Carolina. Following the proposal of Chandrasekar and Deming (1949), all counties (except *Charleston*) are divided into three homogeneous groups based on $0-6$ point index scale which measures the thoroughness of county archives. For more details on the data source and index scaling, readers are referred to Eckberg (2000, pp. 5-9). Data in DRS form are presented in the middle panel of Table 4.

Insert Table 4

In addition to the above two datasets, we consider the **Handloom data**. This new data is from a survey aimed to estimate the undercount in the census of handloom workers (master weavers and labours only) attached to Handloom Industry residing at Gangarampur in South Dinajpur district of state West Bengal, India in 2013. Handloom products have a rich tradition in this state and Handloom Industry occupies a place second only to agriculture in providing livelihood to the people. In the urban area of Gangarampur, there are sixteen wards and out of them two wards (Ward No. 2 and 16) are selected for Post Enumeration Survey to evaluate the coverage in original census (*see* SOSU, 2014). The nature of the

data on these two wards are surprisingly found to be different in terms of recapture proportion. Since, this counting task is meant for the benefit of the workers attached to Handloom industry, a general thought is that the census and PES might be positively related. On the contrary, another issue might underplays in this process is that some people might think one time enrollment is enough. So, if one is counted at the time of census, he/she may be reluctant at the time PES and that yields the underlying ϕ to be less than 1. Thus, in both of possibilities, Chandrasekar-Deming estimate would not be appropriate. Surveyors reported that workers in Ward 16, which is very close to town head-quarter, might be somewhat reluctant to enlist themselves a second time (i.e. at the time of PES). Moreover, most of them are working outside (other districts) and usually come home in particular seasons. That is why, Ward no. 16 results in very low matches compared to Ward no. 2. Moreover, the beliefs of the experts of Textile Directorate of Govt. of West Bengal also drive the idea that the Chandrasekar-Deming estimates (157 and 270 respectively) fails to extract the sizes with precision. However, they expect that Chandrasekar-Deming method yields slight undercount for Ward No. 2 and high overcount for Ward No. 16. Bottom panel of Table 4) presents the DRS data for these two wards.

Insert Table 5

Table 4 presents the list-wise counts and associated matched records from the three data sets mentioned above. Classification results in terms of directional nature of behavioural response for the above three datasets are presented in Table 5 with their corresponding values of the key statistic $\hat{c} = (x_{11}/x_{1\cdot})$. Inference about the directional nature of the

behavioural dependence in the above data sets drawn in literature may not match with the conclusions of the classification strategies proposed in this present article. However, in the light of the current findings, it seems quite plausible that actually the populations *Lilongwe*, *holding index score 6* and *Ward No. 16* are more appropriately classified as *recapture-averse*. These results are quite interesting as they indicate opposite to the conventional idea of *recapture proneness*. Inference for the population *Ward No. 2* drawn from *Rule I* is found to be opposite than that from *Rule II* and *III*. Disagreement is also found in case of the population *holding index score 6*. Indeed, in the conflicting cases, we recommend to go with the classification made by *Rule III* as it is found to be the best in the comparative simulation study.

5 Conclusion

From the extensive literature on capture-recapture data analysis on human population, it is quite clear that list-independence or assumption of *causal independence* does not hold satisfactorily in many instances. As far as homogeneous human population size estimation is concerned, two-sample capture-recapture experiment is appropriate along with M_{tb} modelling. But this model seriously suffers from the non-identifiability problem and analyses in the literature suggest that the availability of the knowledge on nature of behavioural dependency could improve the inference to a great extent. Thus, eliciting such information is crucial. To address this issue, we develop three comparable strategies for classification of the given population (i.e. whether it is *recapture prone* or *averse*) under some mild and realistic assumptions. All the three classification rules are derived based on different threshold

value for $\hat{c}$. The second (*Rule II*) and third (*Rule III*) classification strategies are quite appealing in order to develop more efficient inference in the context of M_{tb}-DRS. Moreover, third strategy is more accurate than second one and it produces nearly 100% success rate except for particular situations with too small recapture probabilities. As we mentioned before, excepting some boundary configurations, third strategy works quite well. Thus, in real life applications, this strategy provides us with a useful tool for classifying populations. Modifications by relaxing the assumptions and extension of this behavioural classification method may be possible for higher order capture occasions (i.e. when $T \geq 3$).

Acknowledgements

Authors gratefully acknowledge the partial financial support by the research fellowship award (Office Order No. DS/JSTK-CC-0020F Dated 11 July, 2013) received by the first author from Indian Statistical Institute, India at the time of his attachment with Indian Statistical Institute, India as a Ph.D. research scholar in Statistics.

References

[1] ChandraSekar, C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. JASA, *44*, 101-115.

[2] Chatterjee, K. and Mukherjee, D. (2016a). An Improved integrated likelihood estimator population size estimation in dual record system. *Statistics and Probability Letters*, *110*, 146-154.

[3] Chatterjee, K. and Mukherjee, D. (2016b). On the Estimation of Homogeneous Population Size from a Complex Dual-record System. *Journal of Statistical Computation and Simulation*, *86*, 3562-3581.

[4] Chatterjee, K. (2016). Some Contributions to the Analysis of Dual-record System for Estimating Human Population Size. Indian Statistical Institute (unpublished doctoral dissertation).

WebLink: http://library.isical.ac.in:8080/xmlui/handle/123456789/6660

[5] Eckberg, D. L. (2000). A capture-recapture approach to the estimation of hidden historical killings, published in *The varieties of homicide and its research: Proceedings of the 1999 meeting of the Homicide Research Working Group, Washington, DC: Federal Bureau of Investigation*. This volume was edited by Blackman, P. H., Leggett, V. L., Olson, B. L., and Jarvis, J. P., pp. 2-10.

[6] El-Khorazaty, M. N. (2000). Dependent dual-record system estimation of number of events: a capture-mark-recapture strategy. *Environmetrics*, *11*, 435-448.

[7] Gosky, R. and Ghosh, S. K. (2011). A Comparative Study of Bayes Estimators of Closed Population Size from Capture-Recapture Data. *Journal of Statistical Theory and Practice*, *5*, 241-260.

[8] Greenfield, C. C. (1975). On the estimation of a missing cell in a 2×2 contingency table. *Journal of Royal Statistical Society A*, **138**, 51-61.

[9] Jarvis, S. N., Lowe, P. J., Avery, A., Levene, S. and Cormack, R. M. (2000). Children are not goldfishâĂŤmark/recapture techniques and their application to injury data. *Injury Prevention*, *6*, 46-50.

[10] Nour, E. S. (1982). On the Estimation of the Total Number of Vital Events with Data from Dual-record Collection Systems. *J. R. Statist. Soc. A*, **145**, 106-116.

[11] Otis, D. L., Burnham, K. P., White, G. C., Anderson, D. R. (1978). Statistical Inference from Capture Data on Closed Animal Populations. *Wildlife Monographs*, *62*, 1-135.

[12] SOSU (2014). Report on the Project "*Survey of Looms and Work sheds in Comprehensive Handloom Development Programme in Dakshin Dinajpur district*âĂİ by Sampling and Official Statistics Unit, Indian Statistical Institute, Commissioned by: Directorate of Textiles, Government of West Bengal, 5th March 2014.

[13] Wang, X., He, C. Z. and Sun, D. (2015). Bayesian Estimation of Population Size via Capture-Recapture Model with Time Variation and Behavioral Response. *Journal of Ecology*, *5*, 1-13.

[14] Wolter, K. M. (1986). Some Coverage Error Models for Census Data. *Journal of American Statistical Association*, *81*, 338-346.

Appendix

Proof of Theorem 1.

$$\begin{aligned} Pr(\text{Recapture Proneness}) &= Pr\left(\frac{3\hat{c}}{2} > \phi_u\right) + \delta_p Pr\left(\frac{3\hat{c}}{2} \leq \phi_u < \sqrt{3}\hat{c}\right) \\ &= 1 - F\left(\frac{2\phi_u}{3}\right) + \delta_p\left[F\left(\frac{2\phi_u}{3}\right) - F\left(\frac{\phi_u}{\sqrt{3}}\right)\right] \\ &= 1 - \left[\delta_p F\left(\frac{\phi_u}{\sqrt{3}}\right) + (1-\delta_p)F\left(\frac{2\phi_u}{3}\right)\right], \end{aligned}$$

where $F(x)$ denoted as the c.d.f of a asymptotically normal variate $\hat{c}$ at x such that $\hat{c}$ has large sample mean c and variance equals to $\sigma^2_{\hat{c}} = V(\hat{c})$. Thus, $F(x) = \Phi\left(\frac{x-c}{\sigma_{\hat{c}}}\right)$, where Φ has its usual meaning, i.e., c.d.f of a std. normal variate.

Now, $Pr(\text{Recapture Aversion}) = Pr(X_p = 0, X_a = 1)$. Therefore,

$$\begin{aligned} Pr(\text{RA}) &= Pr\left(\sqrt{3}\hat{c} < \phi_l, X_p = 0\right) + \delta_a Pr\left(\frac{3\hat{c}}{2} \leq \phi_l \leq \sqrt{3}\hat{c}, X_p = 0\right) \\ &= Pr\left(\sqrt{3}\hat{c} < \phi_l, \sqrt{3}\hat{c} \leq \phi_u\right) + (1-\delta_p)Pr\left(\sqrt{3}\hat{c} < \phi_l, \frac{3\hat{c}}{2} \leq \phi_u < \sqrt{3}\hat{c}\right) \\ &\quad + \delta_a(1-\delta_p)Pr\left(\frac{3\hat{c}}{2} \leq \phi_l \leq \sqrt{3}\hat{c}, \frac{3\hat{c}}{2} \leq \phi_u < \sqrt{3}\hat{c}\right) \\ &\quad + \delta_a Pr\left(\frac{3\hat{c}}{2} \leq \phi_l \leq \sqrt{3}\hat{c}, \sqrt{3}\hat{c} \leq \phi_u\right) \\ &= Pr\left(\sqrt{3}\hat{c} < \phi_l\right) + (1-\delta_p).0 \\ &\quad + \delta_a Pr\left[\frac{\phi_l}{\sqrt{3}} \leq \hat{c} \leq \min\left(\frac{2\phi_l}{3}, \frac{\phi_u}{\sqrt{3}}\right)\right] + \delta_a(1-\delta_p)Pr\left(\frac{\phi_u}{\sqrt{3}} < \hat{c} \leq \frac{2\phi_l}{3}\right). \end{aligned}$$

Now, if ϕ_l and ϕ_u are so chosen such that $\frac{2\phi_l}{3} < \frac{\phi_u}{\sqrt{3}}$, therefore,

$$Pr(\text{RA}) = \delta_a F\left(\frac{2\phi_l}{3}\right) + (1-\delta_a)F\left(\frac{\phi_l}{\sqrt{3}}\right).$$

Hence,

$$\begin{aligned} Pr(\text{CI}) &= Pr(X_p = 0, X_a = 0) \\ &= Pr(X_p = 0) - Pr(X_p = 0, X_a = 1) \\ &= \left[\delta_p F\left(\frac{\phi_u}{\sqrt{3}}\right) + (1-\delta_p) F\left(\frac{2\phi_u}{3}\right)\right] - \left[\delta_a F\left(\frac{2\phi_l}{3}\right) + (1-\delta_a) F\left(\frac{\phi_l}{\sqrt{3}}\right)\right]. \end{aligned}$$

Again, if $\frac{2\phi_l}{3} > \frac{\phi_u}{\sqrt{3}}$ holds,

$$\begin{aligned} Pr(\text{RA}) &= F\left(\frac{\phi_l}{\sqrt{3}}\right) + \delta_a \left[F\left(\frac{\phi_u}{\sqrt{3}}\right) - F\left(\frac{\phi_l}{\sqrt{3}}\right)\right] + \delta_a(1-\delta_p)\left[F\left(\frac{2\phi_l}{3}\right) - F\left(\frac{\phi_u}{\sqrt{3}}\right)\right] \\ &= (1-\delta_a) F\left(\frac{\phi_l}{\sqrt{3}}\right) + \delta_a(1-\delta_p) F\left(\frac{2\phi_l}{3}\right) + \delta_a\delta_p F\left(\frac{\phi_u}{\sqrt{3}}\right) \end{aligned}$$

and similarly,

$$Pr(\text{CI}) = \left[\delta_p(1-\delta_a) F\left(\frac{\phi_u}{\sqrt{3}}\right) + (1-\delta_p) F\left(\frac{2\phi_u}{3}\right)\right] - \left[\delta_a(1-\delta_p) F\left(\frac{2\phi_l}{3}\right) + (1-\delta_a) F\left(\frac{\phi_l}{\sqrt{3}}\right)\right].$$

Table 1: Data Structure in Dual-record System

	List2		
List1	In	out	Total
In	x_{11}	x_{10}	$x_{1\cdot}$
Out	x_{01}	x_{00}	$x_{0\cdot}$
Total	$x_{\cdot 1}$	$x_{\cdot 0}$	$x_{\cdot\cdot} = N$

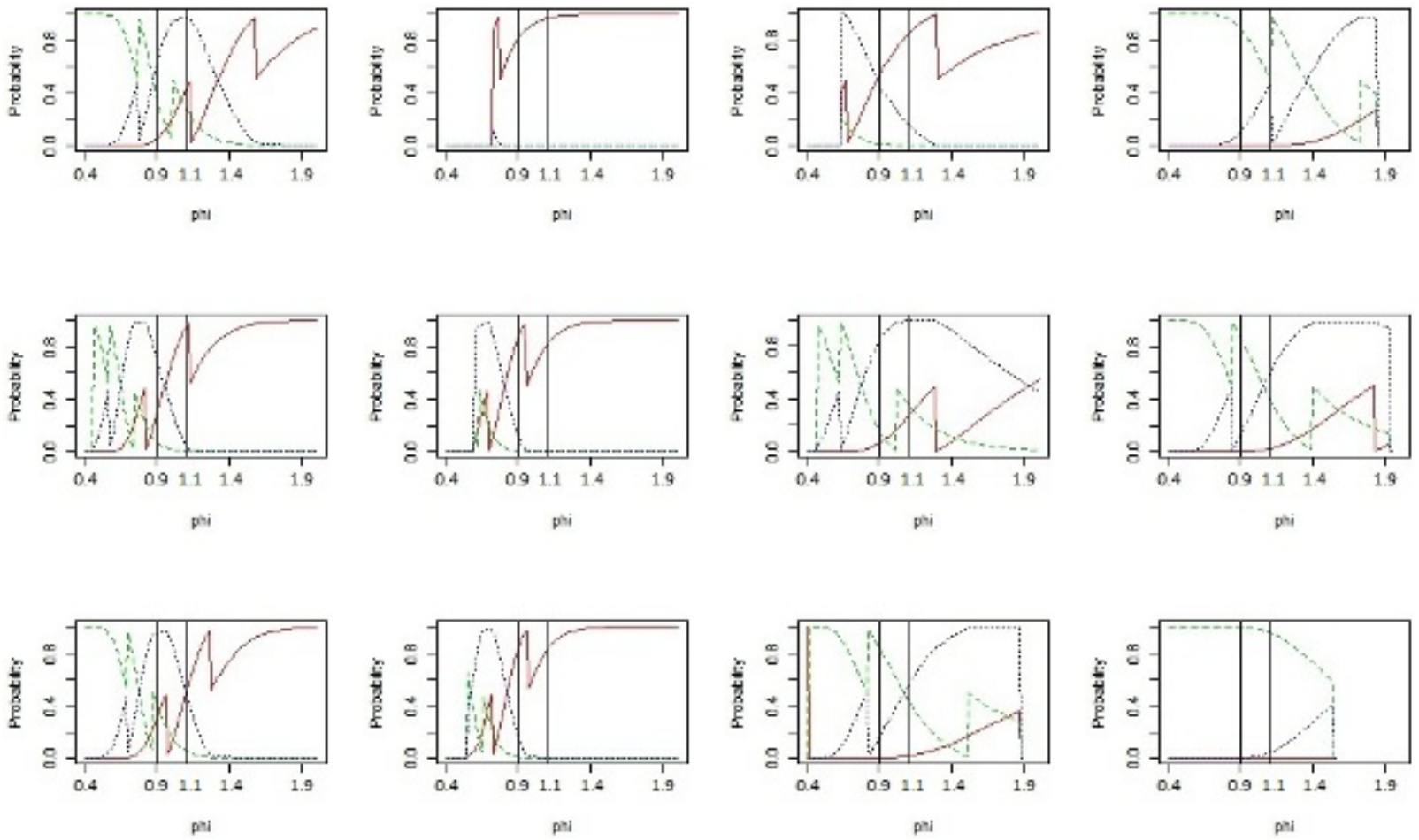

Figure 1: Graphical comparision of three effective probabilities for various ϕ. *Red continuous*, *green dashed* and *blue dotted* lines respectively refer recapture prone, recapture averse and causal independence probabilities.

Table 2: Hypothetical populations considered for simulation study for $N = 1000$

		Recapture Prone		Recapture Averse		Causally Independent	
		$\phi = 1.50$		$\phi = 0.60$		$\phi = 1.00$	
$p_{1\cdot}$	$p_{\cdot 1}$	Population	c	Population	c	Population	c
0.50	0.60	P1	0.72	A1	0.45	I1	0.60
0.70	0.80	P2	0.89	A2	0.67	I2	0.80
0.80	0.70	P3	0.75	A3	0.62	I3	0.70
0.60	0.50	P4	0.58	A4	0.39	I4	0.50
0.55	0.70	P5	0.823	A5	0.538	I5	0.70
0.65	0.80	P6	0.905	A6	0.649	I6	0.80
0.80	0.65	P7	0.611	A7	0.458	I7	0.65
0.70	0.55	P8	0.696	A8	0.574	I8	0.55
0.50	0.70	P9	0.840	A9	0.525	I9	0.70
0.65	0.85	P10	0.996	A10	0.689	I10	0.85
0.85	0.65	P11	0.684	A11	0.590	I11	0.65
0.70	0.50	P12	0.556	A12	0.417	I12	0.50

Table 3: Evaluation of the classification strategy of directional nature of behavioral dependency.

	Rule I	Rule II	Rule III		Rule I	Rule II	Rule III		Rule I	Rule II	Rule III
Population	CCR	CCR	CCR	Population	CCR	CCR	CCR	Population	CCR	CCR	CCR
P1	84.76	99.96	98.18	A1	100.00	100.00	100.00	I1	6.80	69.66	87.36
P2	100.00	100.00	100.00	A2	100.00	100.00	100.00	I2	0.00	0.00	0.00
P3	99.96	100.00	100.00	A3	100.00	100.00	100.00	I3	49.86	0.04	6.72
P4	0.00	0.00	00.94	A4	100.00	100.00	100.00	I4	0.00	0.00	0.14
P5	100.00	100.00	100.00	A5	100.00	99.06	94.00	I5	50.52	0.64	53.46
P6	100.00	100.00	100.00	A6	100.00	99.06	94.00	I6	0.00	0.00	0.00
P7	41.92	99.88	91.32	A7	99.98	78.72	67.82	I7	83.80	49.86	93.96
P8	0.00	1.80	12.40	A8	100.00	100.00	100.00	I8	0.00	2.54	38.24
P9	100.00	100.00	100.00	A9	100.00	99.86	97.92	I9	50.46	0.78	89.48
P10	100.00	100.00	100.00	A10	100.00	99.86	97.92	I10	0.00	0.00	0.00
P11	16.02	98.60	82.46	A11	99.56	42.68	46.58	I11	84.78	49.36	71.14
P12	0.00	0.00	0.02	A12	100.00	100.00	100.00	I12	0.00	0.00	0.06

Table 4: Three real datasets in DRS format which are used for the classification analysis

		Count			
		List 1	List 2	Matched	Total
Data	Populations	$x_{1\cdot}$	$x_{\cdot 1}$	x_{11}	x_0
Malawi Death	Lilongwe	324	216	192	348
	Other Urban Areas	1960	2450	1645	2765
Homicide*	Zero-Two	29	102	23	108
	Three-Five	56	74	42	88
	Six	50	43	32	61
Handloom	Ward No. 2	126	107	85	148
	Ward No. 16	131	103	50	184

*Each of the three populations belong to this datasets are named in terms of holding index score

Table 5: Evaluation of the classification strategy of directional nature of ϕ in M_{tb}

			Nature Classified by		
Data	Populations	$\hat{c}$	Rule I	Rule II	Rule III
Malawi Death	Lilongwe	0.593	Averse	Averse	Averse*
	Other Urban Areas	0.839	Prone	Prone	Prone
Homicide	Zero-Two	0.793	Prone	Prone	Prone
	Three-Five	0.750	Prone	Prone	Prone
	Six	0.630	Averse	Prone	Averse**
Handloom	Ward No. 2	0.657	Averse	Prone	Prone***
	Ward No. 16	0.382	Averse	Averse	Averse

$^{*}\delta = 0.175$; $^{**}\delta = 0.589$; $^{***}\delta = 0.892$